图书情报工作

第 66 卷 第 9 期　2022 年 5 月

文本信息检索中相关性判断的次序效应及其情境因素研究*

■ 王志红[1,2]　曹树金[3]

[1] 清华大学计算机科学与技术系　北京 100084　[2]北京信息科学与技术国家研究中心　北京 100084

[3] 中山大学信息管理学院　广州 510006

摘　要：[目的/意义]基于相关性分值对搜索结果进行列表式线性排列是目前大多数信息检索系统采用的主要方式。已有研究虽然从理论上批判了这种方式的局限性，指出先后检验和阅读的文献之间是相互影响的，但是并未有实证研究进一步调查相关性判断次序效应发生的情境因素。[方法/过程]针对这一问题，采用 2X2X2 被试间实验法，招募 80 名被试，要求每人完成 4 个主题的实验任务，在实验任务完成前后通过问卷调查等方法收集数据，并采用卡方检验、全概率公式背反以及 QQ 等式等多种方法进行数据分析，以探究相关性判断的次序效应以及文献间关系、文献呈现形式作为情境因素所产生的影响。[结果/结论]研究结果表明，文献呈现次序会影响相关性判断，并且发现这种效应会根据文献间关系以及文献呈现形式的不同而不同。具体而言，研究发现相对于具有互补关系的文献，具有互斥关系的文献的呈现次序更容易影响相关性判断；另外，相对于文献全文，文献片段的呈现次序更容易影响相关性判断。本研究虽然存在一定的局限性，但是研究结论对于文本信息检索的理论、实践以及研究方法均具有一定的启示。

关键词：相关性判断　次序效应　文献间关系　文献呈现形式　量子认知

分类号：G250

DOI：10.13266/j.issn.0252-3116.2022.09.013

搜索引擎结果页面(search engine result page，SERP)和加载内容页面(landing page，LP)分别呈现的是检索结果信息项的片段(snippet)和全文，是信息检索系统为用户开展交互活动提供的主要界面。在当前主流的搜索引擎系统中(如百度、谷歌、必应等)，文本信息仍然占据主导性地位，并且 SERP 结果呈现方式仍然以有限数量信息项的顺序式排列为主(一般为 10 条，传统上称之为 ten blue links)，决定了用户在 SERP 上的交互方式主要是以线性的方式检视各个信息项。同时，用户有限的注意力决定了 LP 上信息项的阅读也是以线性的方式展开。其中，判断检索结果(即信息项的片段和全文)的相关性是用户与 SERP 和 LP 交互中的重要环节，这里相关性一般是指信息项及其表征与用户信息需求之间的关系[1]。然而，与信息项的线性交互方式决定了用户在判断当前信息项的相关性时会受到已判断信息项的影响，这种由于信息项的不同呈现次序导致用户相关性评估产生差异的现象被称为次序效应[2]。目前，信息检索系统仍然主要建立在经典的概率排序原则(probability ranking principle，PRP)基础之上，即仅考虑当前单个文献与查询式之间的匹配程度[3]，并没有考虑文献间的相互依赖性[4]及其对信息交互中用户认知的影响，无法纳入用户判断的情境并进行动态建模，也难以实现整个交互过程的累计效用最优。

信息呈现位置或次序如何影响人类认知和行为，在社会科学各个领域开展了较为广泛的研究，例如心理测量研究[5-7]、旅游管理[8-9]、教育学[10]、市场营销学[11-12]、记忆与学习效果[13-14]以及判断与决策领域[15-16]等。在认知心理学领域，该问题更是受到了研究者们的高度关注。近年来，有研究者提出了一种量子认知方法，试图用量子的一些特性(如不可观测性、不兼容性等)来类比和解释人类认知的不确定性以及情境依赖性[17]。在信息检索领域，这一问题也受到了一定的关注。已有研究指出，用户对信息项的相关性

* 本文系国家社会科学基金重大项目"基于特定领域的网络资源知识组织与导航机制研究"(项目编号：12&ZD222)和中国博士后面上资助项目"复杂任务下基于用户相关性反馈的会话搜索优化研究"(项目编号：2021M691823)研究成果之一。

作者简介：王志红，博士后；曹树金，教授，博士，博士生导师，通信作者，E-mail：caosj@mail.sysu.edu.cn。

收稿日期：2021-11-23　修回日期：2022-02-16　本文起止页码：128-140　本文责任编辑：王传清

王志红，曹树金.文本信息检索中相关性判断的次序效应及其情境因素研究[J].图书情报工作,2022,66(9):128－140.

评估会受到各种因素的影响，如判断情境或条件等[18]，其中信息项的呈现次序会对用户认知产生一定的影响[19]，是影响用户相关性判断的重要因素之一[20]。但是，大多数已有研究主要集中在SERP上不同的排序方式(如相关性从高到低或从低到高等)以及不同数量的检索结果对相关性判断[2,18,21]及其一致性[22－23]的影响，或者应用不同相关性判断维度时产生的次序效应[24]等。另外，有少数研究者开始尝试从系统导向的角度，在检索结果排序中建模文献间的依赖关系[25]；李竞飞也尝试使用量子认知方法来进行建模，并区分了4种不同的次序效应，包括比较效应、吸引效应、陌生效应和排斥效应[26]。但是鲜有研究从用户导向的角度，揭示相关性判断中次序效应所发生的不同条件，特别是这些研究都没有将文献间关系和文献呈现形式(包括信息项片段和全文)作为判断情境来考察次序效应的差异。

因此，有必要进一步从用户认知的角度，探索文本信息检索中相关性判断的次序效应，以及在不同文献间关系和不同呈现形式下次序效应的差异，以便深入了解相关性判断的情境性和动态性及其相应的用户认知机制，为用户导向的检索结果排序建模、页面呈现优化提供更细致的理论依据，从而有利于从信息交互过程的角度提升信息检索系统的性能。具体而言，本研究拟采用2X2X2被试间受控实验的方法，旨在探索相关性判断中的次序效应，并考察互补和互斥两种文献间关系以及文献片段和全文两种呈现形式下该效应的差异。其中，互补关系关注的是内容互补这一类关系[27]，指具有互补关系的文献能够提供更多的细节或不同的视角。文献片段和全文两种形式对应的是SERP和LP两种场景中的信息项。其中，文献片段通常包括特定的元数据信息以及文内摘取的部分与查询式匹配的片段。通过这一探索，本研究旨在为理解相关性的动态性及其情境因素提供理论上的洞察，同时为信息检索系统设计提供实践上的启发。

1 研究综述

1.1 搜索结果相关性判断中的位置因素

由于信息需求本质上是一个非常主观且难以清晰界定的概念，因此相关性也具有很强的用户主观导向性，同时也是一个多维度且可测量的概念[28]。在以列表为主要排序方式的信息检索系统中，搜索结果列表中文献的呈现位置或次序会影响到用户相关性的感知和评估[29]。已有研究一方面调查了搜索结果排序列表中结果的位置对用户检验及点击等行为的影响，这些行为可以看成是用户评估文献相关性的一种间接且显式的反映；另一方面则直接调查了文献呈现次序对相关性判断的影响。本研究将主要回顾与第二个问题相关的研究。

在文献呈现次序对用户相关性判断的影响方面，M. Eisenberg和C. Barry通过开展用户实验，发现在相关性升序和降序排列的情况下，用户对文献相关性的评估分别存在高估和低估的现象[2]。在此基础上，M. Huang和H. Wang进一步调查了次序效应及其与文献集大小的关系，发现只有在大小为15到60篇文献集上的相关性判断才会受到次序的显著影响，而在小于15篇或大于60篇时这种影响并不显著[20]。随后，Y. Xu和D. Wang开展了一个两阶段的仿真用户实验，第二个实验阶段的文献集及其次序由第一个实验阶段中的用户行为所决定，研究发现次序效应较小但是是存在的[18]。

由于相关性是一个多维度概念，有研究从判断维度的角度更加深入地探讨和解释了相关性判断中存在的次序效应。例如，P. Bruza和V. Chang[23]提出了一个假设，即相关性判断的每个维度可以被看成是一个希尔伯特空间中的子空间，如果不同文献所引发的用户问题状态通过不同的维度投影到相关性这一整体平面上，这些文献所得到的相关性分值也会不同。通过在众包平台上收集5个查询式有关的数据，从2个不同的维度对各个查询式进行判断，如主题性、可理解性、兴趣等，检验和证实了这一假设。S. Uprety和D. Song将相关性的2个维度扩展到7个维度，并使用查询式日志数据[30]和用户实验[31]等方法检验了相关性维度的次序效应，实验结果证实了相关性维度之间的不可兼容性以及次序效应的存在，并且推测该效应可能与其他因素有关。另外，T. Damessie等调查了次序效应(包括按照降序和文献识别符次序)对相关性判断在不同评价者之间的一致性[21－22]以及判断速度[32]的影响，发现文献呈现次序会影响评价者之间的一致性以及总体评价速度，并且在按文献识别符进行排列时，评价者会花费更少的时间来进行判断且一致性更高。

1.2 文献间关系与文献呈现形式对搜索结果排序的影响

在文献间关系方面，虽然早有研究对信息检索中文献间的独立性假设提出了质疑，但是对文献间关系如何影响用户相关性判断以及如何在结果排序中考虑

这一影响的研究仍然进展甚微。例如,A. Bookstein 指出基于概率决定论模型的方法忽视了文献间交互效应,并认为应该在顺序检索模型(sequence retrieval model)中考虑已检验文献产生的学习效应及其对用户评估后续文献时的影响[33]。D. R. Swanson 提出情报学领域存在几个难以解决的假设,其中之一就是文献之间不是相互独立的,而是处于一个关联网络中,导致相应的判断也处于一个变化性的网络(shift network)之中[34]。更进一步,M. A. Tiamiyu 和 I. Y. Ajiferuke[4]通过从理论上探索文献间关系如何影响用户对单个文献相关性的感知,提出了一个文献交互模型。该模型考虑了3种文献间关系,包括完全没有关系、替代性关系(如重复、冗余等)和互补性关系(即讨论某个特定主题的不同但紧密相关的方面),并建议后续研究应更加深入探索文献间关系及其对相关性感知的影响。

除了文献间关系之外,文献呈现形式也是需要考虑的一个重要因素。已有研究主要考察的是,在相同文献的不同呈现形式下,相关性判断以及所应用的维度会存在差异,但是没有探索不同文献呈现形式下相关性判断次序效应所存在的不同。例如,P. Wang 等[35-36]提出了一个文献选择模型,该模型描述了从书目著录到阅读和引用中的决策过程,并指出分别仅有51.5%和4.9%的检索结果会被阅读和引用,并且不同阶段所采用的判断标准也不同。P. Vakkari 和 N. Hakala 调查了元数据代理和全文两种形式下文献相关性评估的差异,发现在信息搜寻过程中用户对相关性标准的使用与文献呈现形式有关[37]。具体而言,在聚焦前阶段,主题信息在元数据判断比在全文判断中具有更加重要的作用,但是物理特征、信息类型以及用户情境则相反。I. Xie 和 E. Benoit 调查了结果列表和文献全文评估中使用标准的异同,表明在全文评估中使用的标准和相应元素会更多,同时评估所花的时间也会更长[38]。Y. Wang 等应用眼动追踪方法进一步比较了搜索结果列表和单个文献评估的模式,发现两者在固定元素和转移模式方面存在差异,即相比元数据形式,文献全文可以提供更多的信息,从而产生更大的学习效应[39]。

1.3 人类判断中的次序效应及认知心理学解释

在人类判断与决策领域,不少研究者对次序效应产生的心理及认知机制进行了深入的探究和解释,并提出了相应的理论。例如,著名心理学及经济学家 A. Tversky 和 D. Kahneman 在1974年提出了锚定和调整理论用于解释不确定下的判断行为[40]。该理论描述了这样一种现象,即一个人会使用先接收到的数字或信息作为锚点,然后根据随后遇到的信息来调整该锚点。此后,研究者们针对这一现象开展了广泛的研究。例如,R. M. Hogarth 和 H. J. Einhorn 提出了信念调整理论[41],将次序的影响解释为信念调整和更新的结果,后续有研究指出信念调整理论的预测力在涉及极端证据的任务中存在一定的局限[42-43]。

虽然对次序效应发生机制的理论解释存在一定差异,但是研究者们普遍认可次序效应这一现象的存在。一般地,引发这种效应的两个事件被视为具有非兼容性(incompatibility),或称为非兼容性事件。此时,基于集合论的经典概率理论(即不考虑事件出现的次序)或方法在解释不兼容性事件的发生概率时具有一定的局限性,例如难以用传统方法解释的合取谬误或析取谬误[39]。类似的,量子理论中量子的非兼容性等特点也引发了研究者们对经典概率理论或贝叶斯理论的思考,并提出用量子推断模型来解释人类思维的模糊性或非理性人类行为[44-46]。1999年,A. Khrennikov 开始用量子物理有关的理论来解释认知或社会现象,随后他又用量子式模型(quantum-like model)对认知决策及信息加工处理(如囚徒困境或析取谬误)进行建模[47]。一般而言,量子理论的特点在于,它是通过使用希尔伯特空间中的正交子空间进行建模和映射,具有解释不兼容事件的能力,而这种事件间呈现次序的影响无法用不考虑事件出现次序的传统概率理论来解释。

综上可以看出,目前信息检索领域对于相关性判断的次序效应主要集中在呈现次序对相关性判断结果以及判断一致性、速度以及维度等方面的影响,但是少有研究进一步探索这种次序效应发生的情境。特别是,虽然研究者早就意识到文献间关系以及文献呈现形式的重要性,但是尚未有研究将这两个因素作为判断情境探索对次序效应可能产生的影响。与此同时,认知心理学等领域广泛开展了与次序效应有关的研究,为本文探索相关性判断中的次序效应及其情境因素奠定了一定的基础。

2 研究设计与方法

2.1 实验设计与材料选择

本研究一共涉及3个主要变量,包括文献呈现次序、文献间关系和文献呈现形式。本文的基本假设是,文献呈现次序会对相关性判断结果产生影响,并且这种影响受到文献间关系(互补和互斥关系)和文献呈

现形式(文献全文和片段)的作用。首先,对于具有互补或互斥关系的文献而言,不同的文献呈现次序对相关性判断影响的程度可能存在差异;其次,相对于仅呈现文献片段而言,文献全文所产生的次序效应可能会由于信息量的增加而生成更强的学习效应,从而具有不同的次序效应。因此,笔者将采用 2X2X2 的被试间实验设计方法,即一共有 8 个实验小组,所招募的被试将被随机分配到这 8 个实验小组中。

实验要求被试完成所设定的 4 个主题的任务。为了平衡每个被试完成实验的时间,减少来自其他被试的压力,在分发给每位被试的实验材料中,分别有 2 个主题以文献全文和文献片段的形式呈现。由于本文所考虑的互补和互斥关系主要是观点上的互补或互斥,因此争论性议题更适合作为实验主题。通过调查争议性文本理解相关研究中所使用的任务主题,同时考虑到这些议题在本文实验场景中的合适性、流行度等,最终选定的 4 个实验主题见表 1。由于目前所发布的标准数据集均没有考虑文献间关系,因此这 4 个主题的实验材料大多数选自英文版的维基百科。主要理由是因为英文维基百科针对大量争论性议题专门建立了相应的词条,并且词条按照不同的观点来组织,具有非常清晰的主题结构,可以比较容易地为每个主题找到具有互斥和互补关系的子主题。同时,每个子主题需要在维基百科中拥有相应的条目,才能为子主题构建相应的文献对。尽管如此,仍然有些主题无法构建满足条件的实验材料(如没有建立单独的百科条目),因此笔者通过利用谷歌搜索引擎搜索,构建一些与这些子主题有关的新闻作为实验材料。为了节约时间、保证材料的可读性以及减少实验性疲劳带来的影响,所有文献的长度控制在 200－300 字之间。

表 1　所选主题和实验材料

主题	文献关系	文献标题	
主题 1:全球变暖(the causes of global warming)	互斥	森林砍伐(Deforestation)	宇宙射线(Cosmic rays)
	互补	化石燃料(Fossil fuel)	水泥(Cement)
主题 2:转基因食品的安全性(the safety of genetically modified food)	互斥	星联玉米召回(StarLink corn recall)	金色大米(Golden rice)
	互补	等效性(Substantial equivalence)	食物过敏(Food allergy)
主题 3:智能手机的辐射(radiation of smart phone)	互斥	手机辐射不会对人体造成伤害(Radiation From Mobile Phones Are Not Dangerous To Humans)	美国癌症主要机构手机的警告(US cancer boss in mobiles warning)
	互补	蜂窝频率(Cellular frequency)	特定吸收率(Specific absorption rate)
主题 4:视频游戏的影响(the effects of video game)	互斥	视频游戏意味着更少的犯罪(Do video games equal less crime)	消息称 Newtown 杀手受到挪威大屠杀的影响(Newtown shooter motivated by Norway massacre, sources say)
	互补	任天堂手指(Nintendo thumb)	运动控制器(Motion controller)

2.2　实验过程

本研究主要采用便利性抽样的方法来招募被试,这些被试为国内某“双一流”高校信息管理学院选修某一课程的本科生,以尽可能确保被试在阅读能力、英语水平、专业背景等方面的同质性。获得任课教师的同意之后,笔者利用课后时间开展了实验,共有 80 名(注册该课程的学生有 90 名)学生参加了此次实验,时间为 2018 年 11 月。除了 2 名学生没有提交问卷之外,在剩下的 78 名学生(34 名男性,44 名女性)所提交的 312 份(78 人 * 4 份/人)问卷中,剔除 5 份存在缺失值的问卷,最后用于分析的问卷共 307 份。由于实验材料为英文,通过在问卷中设置与英语语言能力有关的问题,发现大部分学生(60 名,76.9%)都通过了全国大学生英语四级考试,表明大部分被试均具有较好的英语阅读技能。同时笔者也挑选出了实验材料中相对较难的词汇,相应给出了中文的翻译,从而进一步避免由于语言障碍引起的实验偏差;同时,在正式实验之前对实验材料的可理解性等开展了小范围前测,以进一步确保所选实验材料的适当性,并且不会引起阅读理解上的问题。

为了削弱实验对象对搜索引擎的信任所产生的结果偏差,本次实验采用纸笔形式开展,并将所有实验材料按顺序排列好之后,根据计划的时间点分别分发给被试。为了保证被试按要求进行实验,一方面,研究人员在实验前着重强调,被试在实验过程中必须严格按照给定的实验材料顺序开展;另一方面,研究人员通过请求两位课程助教的帮助,实时监视整个实验的过程,以防止出现不按要求开展实验的情况。本实验中使用的问卷包括实验前问卷和主调查问卷,实验前问卷用于收集被试个人信息(如性别和英语能力)等,主调查问卷则用于收集被试的相关性判断及其原因。由于量子认知理论的关键假设在于,如果在被调查的第一个

问题之前或两个问题之间存在其他问题，就会改变被试的认知情境，从而产生不准确的测量结果。因此，为了减少实验前问卷对主调查问卷测量题项的结果产生影响，实验前问卷和主调查问卷分别在课前和课后（相隔一个课时，即 45 分钟）开展，这样被试在填写主调查问卷时所处的认知情境就是相同的。在实验开始之前，研究者分发给被试一份实验说明，口头说明实验过程，并紧接着向被试分发实验前问卷。在主实验过程中，分发给被试的实验材料总共包括 4 个任务说明、8 篇提前选定的文献（每个任务 2 篇）和紧随其后的相关性判断题项（4 个刻度，分别为不相关、不太相关、部分相关和相关）。此处使用四点量表，一方面是因为它是研究中应用较多的一种量表，另外已有研究[48]基于类别理论发现，用户在对搜索引擎结果进行分类时平均会产生 3 到 5 个类别，为了便于后续数据分析中对量表进行转换，笔者决定选择四点量表。在每个相关性判断题项的后面，还设置了两个开放式问题，要求被试分别解释对这两篇文献做出判断的原因，即为何他们认为文献是相关或不相关的，以深入理解被试给出相应评估结果的原因。

2.3 数据分析方法

2.3.1 经典概率理论中的全概率公式

在经典概率理论中，事件是以集合的形式来表示的，所有基本事件均由采用布尔逻辑运算符的样本空间所组成。这种概率的集合理论模型是由 A. Kolmogorov 于 1933 年提出并建立的[49]，又被称为 Kolmogorov 模型。该模型中的基本定理是全概率公式。该公式建立在条件概率的基础上，这里条件概率是指在事件 B 出现的条件下事件 A 出现的概率。因此，全概率公式定理是指事件 A 独立出现的概率等于事件 A 在事件 B 出现的所有有限情况下出现概率的总和，如公式（1）所示：

$$p(A) = p'(A) = \sum_k p(B_k) * p(A|B_k) \quad \text{公式(1)}$$

在全概率公式中，最简单的情形为事件 B 只有两种情况，即 B_1 和 B_2。此时，事件 A 的全概率公式如公式(2)所示：

$$p'(A) = p(B_1) * p(A|B_1) + p(B_2) * p(A|B_2) \quad \text{公式(2)}$$

研究表明，基于传统布尔逻辑运算符的全概率公式，在考虑量子力学中的量子概率时会出现背反的情况[50]。这就意味着，当两个事件之间存在量子式干扰（即事件 B 的出现会影响事件 A 出现的概率，或者说两个事件是非独立的或非兼容的）时，事件出现的概率无法用集合理论来表示，即上述等式的左边与右边是不相等的。次序效应，或者概率判断中的合取谬误[51]，就是两个非兼容事件之间存在干扰效应的典型现象。因此，本研究将应用全概率公式来探测相关性判断的次序效应中是否存在量子式干扰效应（quantum-like interference effect），即如果事件 A 独立出现的概率等于事件 A 在事件 B 所有可能情况下出现的全概率，则表明量子式干扰效应不存在，否则存在。

2.3.2 量子等式检验

为了克服经典概率理论在解释不兼容事件时的缺点，Z. Wang 等[17]通过探索 70 个国家级调查中问题对之间的次序效应，以揭示判断与决策中人类认知的量子特性。在该调查数据中，问题对的答案只有两个选项，即是和否，并且这些问题分别以不同的次序进行了调查，通过深入探索这些数据，发现并通过受控实验证明在两种不同次序下所得到的事件列联表中主对角线之和的差值接近于 0。基于这一发现，Z. Wang 等提出了量子等式检验（quantum question test，简称 QQ 等式检验）方法，该方法的具体计算如表 2 和公式（3）所示：

表 2 不同次序下的判断事件

A-B 次序	Br	Bn	B-A 次序	Ar	An
Ar	ArBr	ArBn	Br	BrAr	BrAn
An	AnBr	AnBn	Bn	BnAr	BnAn

$$q值 = (p(ArBr) + p(AnBn)) - (p(BrAr) + P(BnAn)) = (p(AnBr) + p(ArBn)) - (p(BnAr) + P(BrAn)) = 0 \quad \text{公式(3)}$$

在表 2 中，A-B 次序表示事件 A 先出现事件 B 后出现的情况，而 B-A 次序表示相反的情况。Ar 和 Bn 分别表示对于事件 A 和事件 B 其结果为是和否的情况。因此，在两个事件所组成的每种判断次序之下，均存在 4 种不同的情况。例如，在 A-B 次序下存在的 4 种情况分别为 ArBr、ArBn、AnBr 和 AnBn，其概率表示分别为 P(ArBr)、P(ArBn)、P(BrAn)和 P(AnBn)。而 B-A 次序下的 4 种情况则分别为 BrAr、BrAn、BnAr 和 BnAn，其概率表示相应为 P(BrAr)、P(BrAn)、P(BnAr)和 P(BnAn)。QQ 等式就是检验这两种不同次序下所引起的认知情境的改变，或者说第一种次序下所引起的次序效应是否能被第二种次序下的次序效应所抵消。q 值就是这两种情况下抵消程度的测量，因此在次序效应存在的情形下，q 值应该总是接近于 0。满足这一等式的前提假设是：两个事件之间不存在除

了前一个事件之外的其他因素对后一个事件产生的影响[52]。也有研究者提出,QQ 等式是缺少情境性(后一个事件不受前一个事件影响)的充分条件,但不是必要的条件[53]。总体来说,QQ 等式方法是目前应用较多的一种基于量子概率的方法,并且与基于布尔逻辑运算的全概率公式不同,QQ 等式考虑了两个事件取值情况的可能组合,更适合于探索不同情境因素对次序效应的影响这一问题,本研究将在数据分析部分尝试应用该方法对相关性判断的次序效应进行检验。

3 研究结果

3.1 卡方检验结果

3.1.1 总体及各主题下的相关性判断结果分布及差异

在实验中,不同小组的被试被要求利用四点量表为每个主题来评估以不同次序呈现的两篇文献(文献 A 和文献 B),包括文献 A-文献 B 和文献 B-文献 A 两种次序。对于文献 A 而言,以文献 A-文献 B 次序呈现并进行相关性判断的情况,称为非比较场景,即此时文献 A 的相关性判断不会受到其他任何文献的影响;而以文献 B-文献 A 为次序呈现并进行相关性判断时,属于比较场景[54]。对于文献 B 来说,则是相反的,即文献 B-文献 A 为非比较场景而文献 A-文献 B 为比较场景。由于研究获得的数据不满足正态分布(K-S 检验的 p 值小于 0.05,拒绝空假设),因此使用列联表和卡方检验计算比较场景和非比较场景下文献相关性结果的分布及其显著性。表 3 为比较和非比较场景下文献 A 和文献 B 相关性判断分布的列联表,以及比较场景与非比较场景之间的差值。从表 3 可以看出,比较场景与非比较场景的差异主要存在 3 种情况:①相比于非比较场景,在比较场景下判断为相关的数量增加而判断为部分相关的数量减少(对应于总体以及 T1、T4 主题下的文献 A 和 T3 主题下的文献 B),或者判断为部分相关的数量增加而判断为部分不相关的数量减少(对应于 T2 主题下的文献 A),即对某一篇文献而言,另一篇文献对该文献的判断具有增强效应,且这种情况占多数;②相比于非比较场景,在比较场景下判断为部分相关的数量减少,与此同时判断为相关和部分不相关的数量却在增加(对应于总体以及 T2、T4 主题下的文献 B),尤其是部分不相关的数量增加更为明显,即对某一篇文献而言,另一篇文献可能同时对该文献存在减弱效应和增强效应;③比较场景和非比较场景之间的差异并不明显(对应于 T3 主题下的文献 A),即文献之间并没有引起明显的增强或减弱效应。

表 3 比较和非比较场景下文献 A 和文献 B 的相关性判断分布与差异

主题	不同呈现次序及差异	文献 A 的相关性				文献 B 的相关性			
		1	2	3	4	1	2	3	4
T1	文献 A-文献 B/%	0	5	55	40	12.5	27.5	47.5	12.5
	文献 B-文献 A/%	0	2.6	39.5	57.9	7.9	21.1	50	21.1
	△/% =(比较场景 - 非比较场景)	**0**	**-2.4**	**-15.5**	**17.9**	**4.6**	**6.4**	**-2.5**	**-8.6**
T2	文献 A-文献 B/%	5.1	28.2	25.6	41	7.7	28.2	30.8	33.3
	文献 B-文献 A/%	0	13.5	43.2	43.2	2.7	10.8	64.9	21.6
	△/% =(比较场景 - 非比较场景)	**-5.1**	**-14.7**	**17.6**	**2.2**	**5**	**17.4**	**-34.1**	**11.7**
T3	文献 A-文献 B/%	10.3	20.5	28.2	41	0	17.9	38.5	43.6
	文献 B-文献 A/%	7.9	21.1	31.6	39.5	0	23.7	52.6	23.7
	△/% =(比较场景 - 非比较场景)	**-2.4**	**0.6**	**3.4**	**-1.5**	**0**	**-5.8**	**-14.1**	**19.9**
T4	文献 A-文献 B/%	5.1	7.7	43.6	43.6	5.1	38.5	23.1	33.3
	文献 B-文献 A/%	0	5.4	35.1	59.5	10.8	27	37.8	24.3
	△/% =(比较场景 - 非比较场景)	**-5.1**	**-2.3**	**-8.5**	**15.9**	**-5.7**	**11.5**	**-14.7**	**9**
Total	文献 A-文献 B/%	5.1	15.3	38.2	41.4	6.4	28	35	30.6
	文献 B-文献 A/%	2	10.7	37.3	50	5.3	20.7	51.3	22.7
	△/% =(比较场景 - 非比较场景)	**-3.1**	**-4.6**	**-0.9**	**8.6**	**1.1**	**7.3**	**-16.3**	**7.9**

随后,本研究通过卡方检验对这种差异进行显著性检验。总体来看,文献 B 在不同场景下的相关性判断差异具有统计上的显著性($\chi^2 = 8.377$, $p = 0.039 < 0.05$),但是文献 A 在不同场景下的相关性判断差异是不显著的($\chi^2 = 4.568$, $p = 0.206 > 0.05$),并且在两种场景下单元格为 0 的期望数量小于 5。因此,对于文献 B 而言拒绝空假设,而对于文献 A 而言则接受空假设。从不同主题来看,结果显示,只有主题 2 中不同场

景下文献 B 的相关性判断存在统计上显著的差异($\chi^2 = 9.411$, $p = 0.024 < 0.05$)。

3.1.2 不同文献间关系和文献呈现形式的差异

为了检验在不同文献间关系和文献呈现形式的情境下文献相关性判断次序效应是否存在差异,本研究分别针对这两个变量进行了卡方检验,结果分别见表4和表5。其中,A-B 表示先文献 A 后文献 B 这种次序,B-A 则相反。需要注意的一点是,虽然在互斥和互补关系下都用 A 和 B 代表相应的文献对,但是不同文献关系下所使用的实验材料是不一样的,因此不能对同一种次序下(如 B-A 次序下对文献 B 而言)不同文献关系之间进行比较。总体上,对文献 B 而言不同次序对相关性判断的影响是显著的,而对于文献 A 这种影响是不显著的。从文献间关系来看,当文献间关系为互斥时,文献 B 的卡方统计检验结果是显著的($\chi^2 = 8.363$, $p = 0.039 < 0.05$),并且文献 A 双尾检验的 p 值非常接近 0.05($\chi^2 = 7.389$, $p = 0.06$),可认为达到边缘显著。对于具有互补关系的文献而言,文献 A 和 B 均不显著。对于文献呈现形式而言,片段形式的文献 B 在不同场景下的相关性判断差异具有显著性,其他情况均不显著。

表4 次序 * 相关性 * 文献间关系列联表与统计显著性

文献间关系/相关性值		互斥(p=0.06)			互补(p=0.68)			总体(p=0.21)		
		A-B/%	B-A/%	Total/%	A-B/%	B-A/%	Total/%	A-B/%	B-A/%	Total/%
文献 A	1	1.30	0.00	0.60	9.00	4.10	6.60	0.00	2.00	3.60
	2	12.70	2.60	7.70	17.90	18.90	18.40	15.30	10.70	13.00
	3	38.00	35.50	36.80	38.50	39.20	38.80	38.20	37.30	37.80
	4	48.10	61.80	54.80	34.60	37.80	36.20	41.40	50.00	45.60
文献间关系/相关性值		互斥(p=0.04*)			互补(p=0.28)			总体(p=0.04*)		
		A-B/%	B-A/%	Total/%	A-B/%	B-A/%	Total/%	A-B/%	B-A/%	Total/%
文献 B	1	2.50	0.00	1.30	10.30	10.80	10.50	6.40	5.30	5.90
	2	30.40	17.10	23.90	25.60	24.30	25.00	28.00	20.70	24.40
	3	35.40	55.30	45.20	34.60	47.30	40.80	35.00	51.30	43.00
	4	31.60	27.60	29.70	29.50	17.60	23.70	30.60	22.70	26.70

表5 次序 * 相关性 * 文献呈现形式列联表与统计显著性

文献呈现形式/相关性值		全文(p=0.52)			片段(p=0.37)			总体(p=0.21)		
		A-B/%	B-A/%	Total/%	A-B/%	B-A/%	Total/%	A-B/%	B-A/%	Total/%
文献 A	1	5.20	1.30	3.30	5.00	2.70	3.90	5.10	2.00	3.60
	2	7.80	7.90	7.80	22.50	13.50	18.20	15.30	10.70	13.00
	3	39.00	35.50	37.30	37.50	39.20	38.30	38.20	37.30	37.80
	4	48.10	55.30	51.60	35.00	44.60	39.60	41.40	50.00	45.60
文献呈现形式/相关性值		全文(p=0.58)			片段(p=0.05*)			总体(p=0.04*)		
		A-B/%	B-A/%	Total/%	A-B/%	B-A/%	Total/%	A-B/%	B-A/%	Total/%
文献 B	1	5.20	3.90	4.60	7.50	6.80	7.10	6.40	5.30	5.90
	2	20.80	15.80	18.30	35.00	25.70	30.50	28.00	20.70	24.40
	3	40.30	51.30	45.80	30.00	51.40	40.30	35.00	51.30	43.00
	4	33.80	28.90	31.40	27.50	16.20	22.10	30.60	22.70	26.70

3.2 全概率公式背反

在进行全概率公式计算以及 QQ 等式计算之前,需要先进行数据转换,即将相关性的四分类量表分成二分类量表。在选择量表转换方法时,采用的具体操作是将原来相关性值为1、2、3的结果划为不相关一类,而原来相关性值为4的结果划分为相关一类,转换后的数据再用于后续的分析。采用这种转换方式的原因为:首先是为了使转换后的数据尽可能达到平衡。由于文献相关性判断分值分布属于左偏分布,即分值更多地集中于高度相关的区间。正如已有研究中指出的,相关性判断分值集中分布于相关性判断的两个极端(不相关和相关),而较少分布于中间的区间(不太相关和有点相关)[55-56]。其次,本研究更多关注的是文献相关性的评估而不是不相关性的评估,是否或者多大程度上受到其他已判断文献的影响,因此通过这种量表转换方法可以将高度相关性这一类单独区分出

来,更符合这一目的。

在进行数据转换之后,利用全概率公式对比较场景和非比较场景下的概率及其差异进行计算(见表6)。$p(r_i)$和$p(r_j)$分别表示非比较场景下文献 i 和 j 被判断为相关的概率(i 和 j 可以是 A 或 B,当 i 为文献 A 时,j 为文献 B,反之则相反,用星号加以标记以作区分),$p(\bar{r_j})$为非比较场景下文献 j 被判断为不相关的概率,这些概率均是直接从原始数据得到的。$p(r_i|r_j)$是文献 j 被判断为相关的条件下文献 i 被判断为相关的条件概率。相应地,$p(r_i|\bar{r_j})$是文献 j 被判断为不相关的条件下文献 i 被判断为相关的条件概率。$p'(r_i)$是通过公式(2)计算得到的概率。第 1 组到第 4 组文献之间的关系为互斥、第 5－8 组文献之间的关系为互补。从表 6 最后一列可以看出,16 个结果中有 10 个结果的差异超过了 0.1,这一定程度上证实了全概率公式背反的存在,表明文献次序会影响相关性判断。在差异超过 0.1 的所有结果中,有 4 个为互斥关系组、6 个为互补关系组,差异最大的为 0.34。有 3 个为负值,且均来自互补关系组,负值表明相比于非比较场景下,通过全概率公式计算得到的判断为相关的概率有所降低。使用 T 检验方法来分析这两种场景相关性判断的差异是否显著,得到 p 值为 0.003(df 自由度为 15),且皮尔逊相关系数为 0.809(p＝0.000)。这表明,文献次序会显著影响相关性判断。

表 6　比较场景和非比较场景下文献相关性判断的概率及差异

组别_主题	$p(r_i\|r_j)$	$p(r_j)$	$p(r_i\|\bar{r_j})$	$p(\bar{r_j})$	$p'(r_i)$	$p(r_i)$	$\Delta(p'(r_i)-p(r_i))$
G1-4_T1	0.60	0.26	0.64	0.74	0.63	0.40	0.23
G1-4_T1 *	0.00	0.40	1.00	0.60	0.60	0.26	0.34
G1-4_T2	1.00	0.16	0.44	0.84	0.53	0.45	0.08
G1-4_T2 *	0.44	0.45	0.09	0.55	0.25	0.16	0.09
G1-4_T3	0.75	0.21	0.73	0.79	0.74	0.68	0.05
G1-4_T3 *	0.31	0.68	0.50	0.32	0.37	0.21	0.16
G1-4_T4	0.78	0.47	0.40	0.53	0.58	0.40	0.18
G1-4_T4 *	0.75	0.40	0.33	0.60	0.50	0.47	0.03
G5-8_T1	0.67	0.16	0.50	0.84	0.53	0.40	0.13
G5-8_T1 *	0.25	0.40	0.00	0.60	0.10	0.16	－0.06
G5-8_T2	0.33	0.33	0.25	0.67	0.28	0.37	－0.09
G5-8_T2 *	0.71	0.37	0.25	0.63	0.42	0.28	0.14
G5-8_T3	0.00	0.26	0.07	0.74	0.05	0.15	－0.10
G5-8_T3 *	1.00	0.15	0.41	0.85	0.50	0.26	0.24
G5-8_T4	0.00	0.00	0.61	1.00	0.61	0.47	0.14
G5-8_T4 *	0.22	0.47	0.10	0.53	0.16	0.00	0.16

3.3　QQ 等式

对于文献 A-文献 B 次序或文献 B-文献 A 次序,均存在 4 个事件,包括:$A_{rel}B_{rel}$或 $B_{rel}A_{rel}$、$A_{rel}B_{non\text{-}rel}$或 $B_{non\text{-}rel}A_{rel}$、$A_{non\text{-}rel}B_{rel}$或 $B_{rel}A_{non\text{-}rel}$、$A_{non\text{-}rel}B_{non\text{-}rel}$或 $B_{non\text{-}rel}A_{non\text{-}rel}$。$A_{rel}B_{rel}$或 $B_{rel}A_{rel}$表示文献 A 或 B 被判断为相关之后文献 B 或 A 被判断为相关的概率。对原始数据进行转换之后,按照主题、文献间关系和文献呈现形式,对不同次序下每个事件的概率及基于 QQ 等式得到的 q 值进行计算(见表 7)。从表 7 中可以看出,只有主题 3 的 q 值显著小于 0.05,而其他主题的 q 值均大于0.05,表明只有主题 3 通过了 QQ 等式的检验。此外,具有互斥关系或文献呈现形式为片段的情况下的 q 值也小于 0.05,均通过了 QQ 等式检验,说明这些条件下文献相关性判断的次序效应确实可以用量子理论模型来解释。

其他没有通过 QQ 等式检验的情况,原因在于这些情况可能不满足 QQ 等式的前提假设——互惠性原则(the law of reciprocity)[51],即前一个问题是影响回答后一个问题时的情境或状态的唯一因素。尽管在实验设计时尽可能避免在两篇文献之前或者中间,引入除了文献次序之外的其他因素,但是文献是一个非常复杂的对象,可能在阅读并判断前一篇文献时,用户的认知会受到用户自身诸多因素的影响,如背景知识、对该问题的信念或偏好等,从而导致第一篇文献所引起的用户认知变化,对第二篇文献的相关性判断影响较小。此外,还有一个原因可能是在较为真实的实验场景下产生的噪音,会使得这种效应减小以致难以被观察到[57]。当然,从样本量的角度来看,虽然不同主题、

文献间关系以及文献呈现形式之间的样本量是较为平衡的，并且某些情况下已经通过检验，但是仍然需要注意样本量大小可能带来的问题，这也是后续提到的本文局限性之一。

表 7　QQ 等式统计结果及列联表

QQ 值	A-B 次序			B-A 次序		
T1(q = -0.10)	A-B(n = 40)	B_{rel}	B_{non_rel}	B-A(n = 38)	B_{rel}	B_{non_rel}
	A_{rel}	0.050 0	0.350 0	A_{rel}	0.131 6	0.447 4
	A_{non_rel}	0.075 0	0.525 0	A_{non_rel}	0.078 9	0.342 1
T2(q = -0.10)	A-B(n = 39)	B_{rel}	B_{non_rel}	B - A(n = 37)	B_{rel}	B_{non_rel}
	A_{rel}	0.230 8	0.179 5	A_{rel}	0.135 1	0.297 3
	A_{non_rel}	0.102 6	0.487 2	A_{non_rel}	0.081 1	0.486 5
T3(q = 0.01 *)	A-B(n = 39)	B_{rel}	B_{non_rel}	B-A(n = 38)	B_{rel}	B_{non_rel}
	A_{rel}	0.179 5	0.230 8	A_{rel}	0.078 9	0.315 8
	A_{non_rel}	0.256 4	0.333 3	A_{non_rel}	0.157 9	0.447 4
T4(q = -0.10)	A-B(n = 39)	B_{rel}	B_{non_rel}	B-A(n = 37)	B_{rel}	B_{non_rel}
	A_{rel}	0.205 1	0.230 8	A_{rel}	0.189 2	0.405 4
	A_{non_rel}	0.128 2	0.435 9	A_{non_rel}	0.054 1	0.351 4
互斥(q = -0.03 *)	A-B(n = 79)	B_{rel}	B_{non_rel}	B-A(n = 76)	B_{rel}	B_{non_rel}
	A_{rel}	0.177 2	0.303 8	A_{rel}	0.210 5	0.065 8
	A_{non_rel}	0.139 2	0.379 7	A_{non_rel}	0.407 9	0.315 8
互补(q = -0.11)	A-B(n = 157)	B_{rel}	B_{non_rel}	B-A(n = 150)	B_{rel}	B_{non_rel}
	A_{rel}	0.153 8	0.192 3	A_{rel}	0.054 1	0.121 6
	A_{non_rel}	0.141 0	0.512 8	A_{non_rel}	0.324 3	0.500 0
全文(q = -0.15)	A-B(n = 157)	B_{rel}	B_{non_rel}	B-A(n = 150)	B_{rel}	B_{non_rel}
	A_{rel}	0.165 6	0.248 4	A_{rel}	0.133 3	0.366 7
	A_{non_rel}	0.140 1	0.445 9	A_{non_rel}	0.093 3	0.406 7
片段(q = -0.01 *)	A-B(n = 157)	B_{rel}	B_{non_rel}	B-A(n = 150)	B_{rel}	B_{non_rel}
	A_{rel}	0.100 0	0.250 0	A_{rel}	0.094 6	0.067 6
	A_{non_rel}	0.175 0	0.475 0	A_{non_rel}	0.351 4	0.486 5
总体(q = -0.07)	A-B(n = 157)	B_{rel}	B_{non_rel}	B-A(n = 150)	B_{rel}	B_{non_rel}
	A_{rel}	0.165 6	0.248 4	A_{rel}	0.133 3	0.366 7
	A_{non_rel}	0.140 1	0.445 9	A_{non_rel}	0.093 3	0.406 7
	Z = -2.177(p = 0.30)			Z = -4.936(p = 0.000)		

4　讨论与结论

4.1　研究结果讨论

4.1.1　文献呈现次序对相关性判断的影响

（1）定量分析表明文献呈现次序会显著影响相关性判断。综合利用卡方检验、全概率公式背反和 QQ 等式多种定量分析方法发现，即使在只有两篇文献的情况下，相关性判断次序效应也是存在的，并且与被判断的文献本身（包括主题等特征）有关。这一结论与已有研究是一致的，如 M. Huang 和 H. Wang[20] 以及 Y. Xu 和 D. Wang[18] 的研究结果。所不同的是，M. Huang 和 H. Wang 研究认为只有文献数量在特定范围的时候，这种次序效应才会产生。后续研究需要进一步探索不同的文献特征与相关性判断次序效应之间的关联。

（2）开放式问题分析发现被判断的两篇文献之间可能存在比较效应和吸引效应。由于认知能力和时间等的有限性，用户在面对大量相关或部分相关的文献信息时，往往会对不同的文献信息进行比较，以获得对当前任务更加有用的文献信息。因此，相关性判断的次序效应可能与信息交互过程中文献之间的相互比较这一现象有关。例如，有一位被试在解释给出第二篇文献的评估原因时提到“总体看来，比第一篇客观”。还有一位被试在关于第二篇文献的原因陈述中也与前一篇文献进行了比较（“... 对我了解转基因安全给出更客观的原因，而不像上篇给出的例子主观性强”）。此外，定性数据分析还发现次序效应可能还与文献之间的吸引效应或学习效应有关。例如，有被试在第二篇文献的评估原因中提到“结合第一篇的二氧化碳导致全球变暖，本文中提到水泥制作和使用过程会产生大量二氧化碳”，可知第一篇文献的内容为第二篇文献的相关性判断提供了一定的基础，当第一篇文献先出

现时,被试可以从该篇文献中学习,并用于第二篇文献的判断,从而可能会提高第二篇文献判断为相关的概率。

4.1.2 相关性判断次序效应产生的情境

本研究除了调查相关性判断中是否存在次序效应之外,还调查了这种次序效应与文献间关系以及文献呈现形式这两个因素的关系,即相关性判断次序效应所发生的情境。

(1)文献间关系。研究发现,相对于具有互补关系的文献,互斥关系文献的呈现次序更容易影响相关性判断。尽管已有研究从理论上探讨了文献间关系对不同次序下的相关性判断所产生的影响,如 M. A. Tiamiyu 和 I. Y. Ajiferuke[4] 指出,除了考虑单篇文献对信息需求的贡献之外,总体相关性还取决于这篇文献与用户已经检视过的文献之间的关系,包括冗余、重复等替代性关系和互补关系。但是目前尚未发现有研究对这种影响进行检验,因此这一结论无法与已有研究进行比较,且仍然需要进一步验证和探索。

(2)文献呈现形式。研究发现,相对于文献全文,文献片段的呈现次序更容易影响相关性判断。因为文献片段中用于判断相关性的信息更少,用户在进行判断时所形成的信念更不稳定,更容易受到影响。已有研究虽然未具体探讨该问题,但是 I. Xie 和 E. Benoit[38] 证实了不同呈现形式下评估文献相关性所应用的标准和元素存在差异。由于相关性判断是通过应用判断标准并对判断元素进行检视而得到,因此不同文献呈现形式下次序效应的差异,可能一定程度上是由于不同次序的文献所应用的标准和元素不同。

4.2 研究结论

本研究采用便利性抽样的方法招募 80 名被试完成了包含 4 个实验任务的用户实验。通过在实验中控制文献呈现次序、文献间关系以及文献呈现形式,来调查这些因素对用户相关性判断的影响。本研究是首个同时考虑这 3 个因素的用户实验研究,特别是考虑互补和互斥两种文献间关系以及片段和全文两种不同文献呈现形式作为情境因素时,相关性判断次序效应所存在的不同。研究发现,即使在只有两篇文献的情况下,相关性判断的次序效应也是存在的,即文献呈现次序确实会对相关性判断产生影响。此外,本研究也发现,文献间关系和文献呈现形式确实是次序效应的情境因素。具体地,相比于具有互补关系的文献,在具有互斥关系的文献中,相关性判断的次序效应更加明显,同时,相比于文献全文这种呈现形式,对于以文献片段形式来呈现的文献,相关性判断的次序效应更加显著。

4.3 研究启示与局限性

4.3.1 研究价值与启示

已有研究仅仅从理论上探讨了文献间关系对相关性判断次序效应可能存在的影响,本研究首次从用户导向出发,通过实验调查了文献间关系和文献呈现形式作为情境因素,分析了在什么情境下相关性判断次序效应会发生及不同情境产生的差异。

(1)对理论研究的启示。相关性是信息检索领域的一个非常核心且以人为中心的概念。经过半个多世纪的发展,研究者们对相关性的本质仍然没有一个清晰和统一的认识,但是相关性的动态性、多维度性等特征已经受到了广泛的认可。目前,研究者们认为相关性的动态性主要是从信息需求及其问题情境的动态性[58] 和搜索即学习[59] 等理论或视角来加以解释,很少有研究将这种动态性同用户认知情境与信息对象(即文献)之间的交互作用联系起来,并从用户视角来开展相应的研究。缺少对这种交互作用的认识,将会导致无法从理论上深入理解相关性的动态性及其背后的原因或基础规律。因此,后续研究可以从这种视角出发,充分挖掘用户认知与信息对象之间的交互作用,特别是对具有不同关系的信息对象而言,这种交互作用的发生机制及其对用户信息交互过程,特别是对相关性判断产生的影响。

(2)对信息检索系统的启示。目前大部分信息检索系统的基本假设是文献之间是独立的,并基于这一假设分别对文献的相关性进行计算,按照计算所得相关性程度对这些结果进行排序。本研究通过实证研究表明了这一假设的不合理性,并进一步讨论了这一假设不成立的具体条件及其原因,包括文献间互补和互斥关系以及文献呈现形式两个方面作为次序效应产生的情境因素。这一研究结果对于信息检索系统的启示主要有如下几个方面:①在文献间关系对于检索结果排序的启发方面,对于存在互补关系的信息项(特别是文献内细粒度信息内容)而言,虽然具有互补关系的信息项对相关性判断的次序效应并不明显,但是仍需要考虑互补关系中信息项之间的增强或减弱关系,通过信息聚合[60] 的方式将这些信息项关联并综合在一起,实现信息项之间的学习效应或吸引效应,以增强文献集的总体效用,为用户提供更有效的信息检索服务;对于存在互斥关系信息项的排序,不能仅仅考虑与检索需求的匹配性程度,以及与用户已有点击历史的相似程度来返回检索结果,否则会导致用户接收的信息窄化并产生典型的信息茧房[61],而是应该考虑这种具有互斥关系的信息项对用户认知产生的影响,如通过交

叉排列的方式依次或平行地呈现具有互斥关系的信息项，使搜索结果多样化[62]。一方面，这样可以便于用户从不同的角度对这些信息项进行比较；另一方面也可以增加搜索结果中排在靠后文献的边际效用值[63]，提高用户对搜索结果集的满意度。②在文献呈现形式方面，研究发现文献片段更容易引起相关性判断的次序效应，因此如何用文献片段的方式更充分地描述和揭示文献的内容，让用户形成更加稳定的信念也是信息检索系统应该考虑的另一个问题。

（3）对研究方法的启示。本研究应用量子认知心理学中的方法对数据进行统计和分析，更加全面、综合地检验了次序效应的存在以及发生的情境，认为基于量子理论的数理分析与统计方法在信息检索研究中的应用是非常具有前景的。目前，与量子有关的现象已经被广泛地证实和发现，并且研究者认为将量子理论应用于解释人类认知行为具有一定的合理性和适用性[64]。信息检索早已经从系统导向转向了认知导向和情境导向，这两种导向下的信息检索必然涉及大量人类认知和情境问题，尤其是信息的相关性判断和选择。例如，量子理论中的基本原理之一——互补性原理，与信息检索中相关性判断次序效应不谋而合，即互补性原理认为测量必须按照次序进行并且第一次测量产生的情境会影响接下来测量的结果，从而产生测量的次序效应。

4.3.2 研究局限性

本研究采用便利性抽样的方式招募某“双一流”高校的学生作为被试，样本覆盖范围和样本量比较有限，此为本研究的局限性之一。研究的局限性之二在于实验材料的选择方面。尽管笔者从多个方面考虑了实验材料选择的合理性，但是由于缺少考虑文献间关系的标准化文献材料，如 TREC 会议所提供的标准化评测过程和文献材料，因此笔者在选择文献材料时仍然不乏一定的武断性和主观性。在缺少标准化、可比较的文献材料作为实验材料的情况下，这种文献选择的主观性是信息检索领域实验研究中难以克服的一个常见性问题。因此，研究者需要在尽可能详尽地考虑文献有关的各种变量对实验结果产生的影响，控制这些因素对实验结果的干扰。未来研究可从如下几个方面来进一步开展：①后续研究可进一步调查多模态（如图片、视频等）信息检索中的次序效应问题，包括单个模态下以及不同模态间信息对象的相关性判断次序效应问题；②后续研究还可以考虑不同垂直领域下（如百科、博客、新闻等）信息对象之间相关性判断的次序效应问题；③后续研究可以考察不同的文献特征（如不同的文献间关系）以及不同的用户特征（如用户阅读能力、认知风格、与主题有关的信念等）情境下，相关性判断次序效应的差异。

致谢：非常感谢芬兰坦佩雷大学信息检索研究小组 Jaana Keikalainen 教授和 Vakkari Pertti 教授在开展实验和结果分析中所提供的指导与帮助。

参考文献：

[1] SARACEVIC T. Relevance: a review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance[J]. Journal of the American Society for Information Science and Technology, 2007, 58(13): 1915－1933.

[2] EISENBERG M, BARRY C. Order effects: a study of the possible influence of presentation order on user judgments of document relevance[J]. Journal of the American Society for Information Science, 1988, 39(5): 293－300.

[3] PANG L, AI Q, XU J. Beyond probability ranking principle: modeling the dependencies among documents[C]//Proceedings of the 14th ACM international conference on Web search and data mining. New York: ACM,2021: 1137－1140.

[4] TIAMIYU M A, AJIFERUKE I Y. A total relevance and document interaction effects model for the evaluation of information retrieval processes[J]. Information processing & management, 1988, 24(4): 391－404.

[5] 王思琦，郭金云. 公共服务满意度测量的问题顺序效应：来自一项嵌入性调查实验的证据[J]. 公共管理评论，2020，2(1)：92－115.

[6] TOURANGEAU R, RASINSKI K A. Cognitive processes underlying context effects in attitude measurement[J]. Psychological bulletin, 1988, 103(3): 299.

[7] 孙玉，卢淳，司继伟，等. 任务呈现方式、任务呈现顺序影响青少年算术策略运用的年龄差异[J]. 心理研究，2017，10(1)：39－49.

[8] 胥兴安，王立磊，高峰强. 旅游广告与网络负面口碑对目的地形象的影响——次序效应和交互效应的实验检验[J]. 旅游学刊，2017，32(12)：37－48.

[9] ERT E, FLEISCHER A. Mere position effect in booking hotels online[J]. Journal of travel research, 2016, 55(3): 311－321.

[10] 霍航丰. 项目呈现顺序对学习判断的影响[D]. 杭州：浙江师范大学，2017.

[11] PURNAWIRAWAN N, DE PELSMACKER P, DENS N. Balance and sequence in online reviews: how perceived usefulness affects attitudes and intentions[J]. Journal of interactive marketing, 2012, 26(4): 244－255.

[12] 谢梦娇. 追加式评论中顺序效应与消费者购买意愿的关系研究[D]. 北京：北京邮电大学，2018.

[13] 许智慧. 不同难度教学视频中陈述性知识和程序性知识的呈现顺序对学习的影响[D]. 武汉：华中师范大学，2019.

[14] 王家慰，邓泳珊，邢强. 学习顺序和材料相似性对听觉类别学习的影响[J]. 心理与行为研究，2019，17(6)：750－755.

[15] BERGUS G R, LEVIN I P, ELSTEIN A S. Presenting risks and benefits to patients: the effect of information order on decision making[J]. Journal of general internal medicine, 2002, 17(8): 612

－617.

[16] 杨文婷. 信息呈现对个体投资决策的影响[D]. 杭州:浙江大学, 2019.

[17] WANG Z, SOLLOWAY T, SHIFFRIN R M, et al. Context effects produced by question orders reveal quantum nature of human judgments[J]. Proceedings of the National Academy of Sciences, 2014, 111(26): 9431－9436.

[18] CHU H. Factors affecting relevance judgment: a report from TREC legal track[J]. Journal of documentation, 2011, 67(2):264－278.

[19] XU Y, WANG D. Order effect in relevance judgment[J]. Journal of the American Society for Information Science and Technology, 2008, 59(8): 1264－1275.

[20] GOLDMAN E. Search engine bias and the demise of search engine utopianism [M]//Web Search. Berlin: Springer, 2008: 121－133.

[21] HUANG M, WANG H. The influence of document presentation order and number of documents judged on users' judgments of relevance[J]. Journal of the American Society for Information Science and Technology, 2004, 55(11): 970－979.

[22] DAMESSIE T T, SCHOLER F, JÄRVELIN K, et al. The effect of document order and topic difficulty on assessor agreement[C]//Proceedings of the 2016 ACM international conference on the theory of information retrieval. New York: ACM, 2016: 73－76.

[23] DAMESSIE T T, CULPEPPER J S, KIM J, et al. Presentation ordering effects on assessor agreement[C]//Proceedings of the 27th ACM international conference on information and knowledge management. New York: ACM, 2018: 723－732.

[24] BRUZA P, CHANG V. Perceptions of document relevance[J]. Frontiers in psychology, 2014, 5(612): 1－8.

[25] ZUCCON G, AZZOPARDI L. Using the quantum probability ranking principle to rank interdependent documents[C]//European conference on information retrieval. Berlin: Springer, 2010: 357－369.

[26] 李竞飞. 面向探索式信息检索的量子交互模型[D]. 天津: 天津大学, 2017.

[27] MA Q, NADAMOTO A, TANAKA K. Complementary information retrieval for cross-media news content[J]. Information systems, 2006, 31(7): 659－678.

[28] SCHAMBER L, EISENBERG M B, NILAN M S. A re-examination of relevance: toward a dynamic, situational definition[J]. Information processing & management, 1990, 26(6): 755－776.

[29] SCHAMBER L. Relevance and information behavior[J]. Annual review of information science and technology, 1994, 29(1): 3－48.

[30] UPRETY S, SONG D. Investigating order effects in multidimensional relevance judgment using query logs[C]//Proceedings of the 2018 ACM SIGIR international conference on theory of information retrieval. New York: ACM,2018: 191－194.

[31] LI J, ZHANG P, SONG D, et al. Understanding an enriched multidimensional user relevance model by analyzing query logs[J]. Journal of the Association for Information Science and Technology, 2017, 68(12): 2743－2754.

[32] DAMESSIE T T, SCHOLER F, CULPEPPER J S. The influence of topic difficulty, relevance level, and document ordering on relevance judging[C]//Proceedings of the 21st Australasian document computing symposium. New York: ACM,2016: 41－48.

[33] BOOKSTEIN A. Information retrieval: a sequential learning process[J]. Journal of the American Society for information Science, 1983, 34(5): 331－342.

[34] SWANSON D R. Historical note: information retrieval and the future of an illusion[J]. Journal of the American Society for Information Science, 1988, 39(2): 92－98.

[35] WANG P, SOERGEL D. A cognitive model of document use during a research project. Study I. Document selection[J]. Journal of the American Society for Information Science, 1998, 49(2): 115－133.

[36] WANG P, DOMAS WHITE M. A cognitive model of document use during a research project. Study II. Decisions at the reading and citing stages[J]. Journal of the American Society for Information Science, 1999, 50(2): 98－114.

[37] VAKKARI P, HAKALA N. Changes in relevance criteria and problem stages in task performance[J]. Journal of documentation, 2000, 56(5):.540－562.

[38] XIE I, BENOIT E. Search result list evaluation versus document evaluation: similarities and differences[J]. Journal of documentation, 2013, 69(1): 49－80.

[39] WANG Y, XIE I, LEE S. Explore eye movement patterns in search result evaluation and individual document evaluation[J]. Proceedings of the Association for Information Science and Technology, 2015, 52(1): 1－4.

[40] TVERSKY A, KAHNEMAN D. Judgment under uncertainty: Heuristics and biases[J]. Science, 1974, 185(4157): 1124－1131.

[41] HOGARTH R M, EINHORN H J. Order effects in belief updating: the belief-adjustment model[J]. Cognitive psychology, 1992, 24(1): 1－55.

[42] TRUEBLOOD J, BUSEMEYER J. A comparison of the belief-adjustment model and the quantum inference model as explanations of order effects in human inference[C]//Proceedings of the Annual Meeting of the Cognitive Science Society. Seattle: Cognitive Science Society,2010, 32(32): 1166－1171.

[43] TRUEBLOOD J S, BUSEMEYER J R. A quantum probability account of order effects in inference[J]. Cognitive science, 2011, 35(8): 1518－1552.

[44] BUSEMEYER J R, WANG Z, KHRENNIKOV A, et al. Applying quantum principles to psychology[J]. Physica scripta, 2014(T163): 014007.

[45] BUSEMEYER J R, KVAM P D, PLESKAC T J. Markov versus quantum dynamic models of belief change during evidence monitoring[J]. Scientific reports, 2019, 9(1): 1－10.

[46] AERTS D, GABORA L, SOZZO S. Concepts and their dynamics: a quantum-theoretic modeling of human thought[J]. Topics in cognitive science, 2013, 5(4): 737－772.

[47] KHRENNIKOV A. Quantum-like model of cognitive decision making and information processing[J]. Biosystems, 2009, 95(3): 179－187.

[48] ZHITOMIRSKY-GEFFET M, BAR-ILAN J, LEVENE M. Categorical relevance judgment[J]. Journal of the Association for Information Science and Technology, 2018, 69(9): 1084－1094.

[49] KHRENNIKOV A. The principle of supplementarity: a contextual

probabilistic viewpoint to complementarity, the interference of probabilities and incompatibility of variables in quantum mechanics [J]. Foundations of physics, 2005, 35(10): 1655 – 1693.

[50] KHRENNIKOV A. Quantum-like modeling of cognition[J]. Frontiers in physics, 2015, 3(77): 1 – 18.

[51] TVERSKY A, KAHNEMAN D. Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment[J]. Psychological review, 1983, 90(4): 293.

[52] WANG Z, BUSEMEYER J R. A quantum question order model supported by empirical tests of an a priori and precise prediction [J]. Topics in cognitive science, 2013, 5(4): 689 – 710.

[53] DZHAFAROV E N, ZHANG R, KUJALA J. Is there contextuality in behavioural and social systems? [J]. Philosophical transactions of the Royal Society A: mathematical, physical and engineering sciences, 2016, 374(2058): 20150099.

[54] MOORE D W. Measuring new types of question-order effects: additive and subtractive[J]. The public opinion quarterly, 2002, 66 (1): 80 – 91.

[55] GREISDORF H. Relevance: an interdisciplinary and information science perspective[J]. Informing science, 2000, 3(2): 67 – 72.

[56] JANES J W. Relevance judgments and the incremental presentation of document representations[J]. Information processing & management, 1991, 27(6): 629 – 646.

[57] YEARSLEY J M, BUSEMEYER J R. Quantum cognition and decision theories: a tutorial[J]. Journal of mathematical psychology, 2016, 74: 99 – 116.

[58] BELKIN N J. Anomalous states of knowledge as a basis of information retrieval[J]. The Canadian journal of information science, 1980, 5(1): 133 – 143.

[59] VAKKARI P. Searching as learning: a systematization based on literature[J]. Journal of information science, 2016, 42(1): 7 – 18.

[60] 曹树金, 李洁娜, 王志红. 面向网络信息资源聚合搜索的细粒度聚合单元元数据研究[J]. 中国图书馆学报, 2017, 43(4): 74 – 92.

[61] 王益成,王萍,王美月. 基于 SVM 的网络信息茧房层次敏感影响因素识别研究[J]. 情报资料工作, 2019,40(6):90 – 97.

[62] DZUNG H, LUO S. Search result diversification in resource selection for federated search[C]// International ACM SIGIR conference on research & development in information retrieval. New York: ACM, 2013: 612 – 622.

[63] YU H, JATOWT A, BLANCO R, et al. Decoding multi-click search behavior based on marginal utility[J]. Information retrieval, 2017, 20(1): 25 – 52.

[64] BUSEMEYER J R, WANG Z. What is quantum cognition, and how is it applied to psychology? [J]. Current directions in psychological science, 2015, 24(3):163 – 169.

作者贡献说明:
王志红:设计实验方案,开展实验,起草论文;
曹树金:修改实验方案,修改论文。

Research on Order Effect in Relevance Judgment and Its Contextual Factors in Textual Information Retrieval

Wang Zhihong[1,2]　Cao Shujin[3]
[1] Department of Computer Science and Technology, Tsinghua University, Beijing 100084
[2] Beijing National Research Center for Information Science and Technology, Beijing 100084
[3] School of Information Management, Sun Yat-sen University, Guangzhou 510006

Abstract: [**Purpose/Significance**] The main approach adopted in most information retrieval systems currently is to present the search results list linearly based on the relevance score. Although existing researches have theoretically criticized the limitations of this approach, proposing that there is mutual influence between the document previously examined and read, there is no empirical study to further investigate the contextual factors affecting the order effect of relevance judgment. [**Method/Process**] In response to this question, this paper adopted the 2X2X2 inter-subject experiment method by recruiting 80 subjects. Each subject was required to complete 4 experimental tasks with different topics. Data was collected by questionnaires before and after each experimental task was completed. In order to explore the order effect of relevance judgment and the influence of inter-document relationships as well as the document presentation types as contextual factors, the collected data was analyzed using various methods, such as chi-square test, law of total probability violation, QQ equation, etc. [**Result/Conclusion**] The research results confirm that the judgment of relevance can be affected by the presentation order of the documents and this effect is different according to the relationships between documents and document presentation types. Specifically, compared to documents with complementary relationships, the order of documents with contradictory relationship is more likely to influence the relevance judgment. In addition, compared with the full text, the order of the presentation of literature snippets is more likely to affect the relevance judgment. Although this paper has its limitations, the research conclusions obtained have certain theoretical, practical and methodological enlightenment.

Keywords: relevance judgment　order effect　inter-document relationship　document presentation type　quantum cognition